CREATION DE CORPUS REFLECHIS POUR L'ENSEIGNEMENT DE LA PRODUCTION ECRITE. L'EXEMPLE DE L'ALLEMAND, LANGUE ETRANGERE.

Eva Schaeffer-Lacroix
evalacroix@free.fr
Syled
Paris 3
France

Abstract : Corpora are considered as resources helping to improve teaching and learning practice. However, there is a lack of pedagogical resources suited for foreign language learners at a low or intermediate level. After a description of research objectives and corpora types, I will present tools helping to create and explore German corpora, always keeping in mind the context of language teaching.

1. Introduction

O'Keeffe, McCarthy et Carter (2007) incitent à une utilisation de corpus par des enseignants et apprenants de langue étrangère. Ces ressources textuelles, destinées à être explorées afin d'étudier la langue, seraient porteuses d'une évolution des pratiques et des attitudes d'apprentissage. Braun et Chambers (2006) précisent pourquoi les recherches dans ce domaine n'ont pas encore eu de répercussions sensibles sur la pratique : les corpus utilisables pour l'apprentissage des langues, gratuits et faciles d'accès, sont encore très difficiles à trouver. Lüdeling et Walter (2009 : 7) confirment ce constat pour les ressources adaptées à des apprenants de niveau peu avancé à intermédiaire (A2/B1) en allemand. La rareté de ressources en langue allemande, compatibles avec des scénarios constructivistes, susceptibles de soutenir la prise en compte de normes linguistiques et discursives, a motivé mon choix de créer des corpus de petite taille pour un projet de rédaction dans le domaine du tourisme. J'ai proposé ce projet à des apprenants francophones en quatrième année d'allemand, ayant entre 14 et 15 ans, dans dans le cadre de leurs cours réglementaires au collège. Il s'agit de corpus réfléchis, donc de corpus rassemblant des textes entiers qui représentent un genre textuel spécifique (Rastier, 2004). J'ai conçu ces corpus comme outils d'aide à la rédaction en LE (Langue Étrangère), complétant les aides existantes, à savoir les dictionnaires et les traducteurs électroniques.

Dans mon article, je préciserai d'abord les objectifs de l'utilisation de corpus et les conséquences qui en découlent pour leurs caractéristiques spécifiques. J'expliquerai ensuite en quoi les corpus allemands existants auxquels j'ai eu accès en sont partiellement dépourvus. Je décrirai trois des corpus réfléchis que j'ai créés. Je confronterai les performances de l'un d'eux aux performances de trois corpus allemands existants. Je terminerai par la description des étapes de création de corpus réfléchis de petite taille visant à être accessibles aux linguistes et didacticiens qui ne sont pas spécialisés en informatique. Il sera question d'outils de création de corpus permettant de rassembler, d'annoter et d'interroger des données trouvées sur la Toile. J'adopterai un point de vue large permettant de transférer ma démarche à d'autres langues. Les constantes que j'ai retenues sont les corpus réfléchis, le domaine de la production écrite et le niveau d'apprentissage "peu avancé à intermédiaire".

2. Objectifs de l'utilisation de corpus

Pourquoi avoir recours à un corpus ? Le tableau ci-dessous, inspiré de Mayaffre (2005), décrit le lien entre le positionnement des linguistes de corpus par rapport à la théorie et le choix du type de corpus. Il décrit, de façon schématisée, les approches méthodologiques et les objectifs en fonction de ce choix.

Tableau 1 – Postures de recherche vis-à-vis des corpus

Type de corpus	Corpus lexicographiques (extraits de textes attestés).	Corpus réfléchis (textes entiers attestés).	Corpus phrastiques (phrases formulées par des linguistes).
Théorie	Corpus "sans théorie".	Corpus utilisé pour créer la théorie.	La théorie préexiste au corpus.
	Parole.	Parole/langue.	Langue.
	Linguistique de corpus ; linguistique descriptive.	Linguistique sur corpus.	Linguistique théorique.
Méthodologie	Empirisme	Empirisme/théorie.	Théorie.
	Induction.	Induction/déduction.	Déduction.
Objectif	Le corpus sert d'apport pour construire le savoir.		Le corpus sert à valider et à illustrer des hypothèses et théories qui existent déjà.

2.1. La description de la langue

L'objectif général de l'utilisation d'un corpus est l'étude de la langue. Si les descriptions partent d'énoncés attestés, sans s'appuyer sur un cadre théorique, elles risquent d'avoir peu de valeur scientifique. Si le corpus est abordé dans le but d'illustrer des théories linguistiques existantes, certains faits attestés peuvent rester ignorés.

2.2. L'étude de l'espace des normes linguistiques

Bommier-Pincemin (1999 : 421) propose d'appliquer aux corpus le terme d'"espace discursif". Selon Rastier (2004), "la linguistique de corpus peut offrir les moyens théorique et technique d'étudier l'espace des normes". Il considère cet espace comme le chaînon manquant entre langue et parole. Grâce à la juxtaposition de multiples actes d'énonciation individuels attestés, les corpus sont l'expression d'un schéma général de l'énonciation (Maingueneau, 2008 : 10). Cela leur confère un statut de norme déterminée par les genres et les discours. Les normes de l'orthographe, par exemple, s'en trouvent relativisées : elles peuvent varier en fonction du genre textuel. Les genres liés à la publication en ligne en témoignent tout particulièrement.

2.3. L'apprentissage de la production écrite

Les corpus peuvent être le support matériel de tâches dans le cadre de l'enseignement-apprentissage d'une langue, telles que la recherche d'informations ou l'observation d'un phénomène linguistique. Appliquée à la didactique de la production écrite en LE, l'étude de l'espace des normes permet de repenser le statut de notions comme "règle" et "erreur". Leur pertinence peut être interrogée grâce à l'observation d'énoncés attestés. Les représentations des apprenants et des enseignants concernant la langue étrangère peuvent ainsi être mises à l'épreuve et, le cas échéant, être modifiées.

Teubert (2006) rappelle qu'un nouveau texte est toujours joint à un discours existant. Ce point de vue rapproche les corpus et l'apprentissage de la production écrite. Les corpus réfléchis favorisent une approche discursive et pragmatique, puisqu'on dispose de textes entiers dont on connaît l'usage social. Il est possible d'inciter les apprenants à observer quelles sections caractérisent les textes représentant un même genre textuel (entrée "top down"). Les lignes de concordance facilitent l'observation des caractéristiques linguistiques

des sections et des textes entiers à un niveau "bottom up". On peut s'intéresser à la valence verbale, aux relations collocationnelles d'un lexème, aux temps et modes, etc.

3. Quels corpus pour un projet d'écriture en LE ?

3.1. Types de corpus

Dans le cadre d'un projet d'écriture qui s'intéresse aux genres textuels, j'écarte les corpus équilibrés (Bommier-Pincemin, 1999 : 417), cherchant à représenter "la globalité" des données dont on dispose concernant une langue, et les corpus d'échantillons (Habert, Nazarenko et Salem, 1997 : 144). Je retiens le corpus réfléchi, contenant des documents entiers qui représentent tous un même genre textuel (Rastier, 2004). Ceci permet de travailler avec des données qui ont des caractéristiques linguistiques et discursives comparables. Je rapproche les corpus réfléchis des corpus spécialisés qui visent à offrir la représentation linguistique d'une situation de communication, d'un domaine, d'un profil de locuteur ou d'une langue de spécialité (Habert, Nazarenko et Salem, 1997 : 144 ; Bowker et Pearson, 2002 : 12). La condition de rassembler des textes entiers est constitutive pour les corpus réfléchis. Elle ne l'est pas forcément pour les corpus spécialisés.

Les corpus dits "comparables" et les corpus parallèles, juxtaposant du texte en deux langues ou plus, sont souvent explorés par les traducteurs et les lexicologues. Ils pourraient également trouver leur place au sein d'un projet d'écriture en langue étrangère (Narita, Kurokawa et Utsuro, 2003). Je n'ai pas eu accès à des corpus mettant en parallèle des textes français et allemands et représentant un des genres textuels que j'ai choisis pour ma recherche (trois genres textuels du domaine du tourisme, critiques de film, annonces pour trouver un correspondant). Le corpus parallèle *English-German Translation Corpus* comporte, entre autres, des données du domaine du tourisme. Toutefois, il juxtapose l'allemand et l'anglais, et son utilisation a certains inconvénients que je détaillerai plus loin.

3.2. Corpus allemands existants

Peu de ressources actuellement accessibles en ligne couvrent des domaines pouvant convenir pour des projets d'apprentissage de LE à un niveau peu avancé à intermédiaire. L'ensemble de corpus *DeReKo* (Deutsches ReferenzKorpus) [Corpus de référence allemand] donne l'occasion d'interroger un choix de 89 de ses corpus à l'aide de l'outil de recherche *COSMAS II* (Corpus Search, Management and Analysis System). Le contenu de certains des corpus paraît pertinent pour un public d'apprentis-scripteurs. En 2008, environ 500 millions de mots sont classés sous le thème "Loisirs, divertissement" (Kupietz et Keibel, 2008). Le corpus *MK1* (Mannheimer Korpus 1) contient, entre autres, 101.562 mots représentant les genres textuels "biographie, autobiographie, journal intime". Toutefois, les textes de *MK1* ont été rédigés entre 1950 et 1967. Je m'intéresse à des textes plus récents. S'y ajoute à un deuxième inconvénient : l'utilisation de *DeReKo* peut actuellement paraître difficile à des non-spécialistes en linguistique de corpus. Le développement d'une interface plus facile d'emploi est en élaboration. *Cosmas II* reste donc une perspective intéressante.

3.3. Corpus créés pour le projet

La rareté de corpus allemands ayant un contenu et une interface qui semblent pouvoir convenir à un scénario d'apprentissage de la production de tâches écrites en allemand justifie mon choix de créer des corpus en fonction des besoins spécifiques. J'ai constitué trois corpus non annotés qui sont de très petite taille (entre 3.400 et 9.800 mots). Ceci m'a permis de travailler avec des données relativement "maîtrisées". J'ai pu vérifier manuellement le contenu et l'orthographe. Pour chacun de ces corpus, j'ai créé un fichier contenant le texte sous forme de lignes de concordance. Pour ce faire, je me suis servie du concordancier *Text-based concordances (v2)* de Cobb (nd) (cf. 4.4.).

Rastier (2004) recommande l'indication d'une triple information dans l'en-tête des corpus, à savoir le discours, le champ générique et le genre textuel. Les textes des trois corpus du projet de rédaction peuvent être décrits selon ces critères. Le champ générique est celui du tourisme. Les genres et discours représentés dans ce projet sont les "Informations pratiques" (discours informationnel), la "Visite guidée pour jeunes" (discours publicitaire) et le "Commentaire dans un livre d'or électronique d'un musée" (discours de type épistolaire). Ces trois genres textuels peuvent être repérés sur un grand nombre de sites Internet de musées et d'établissements culturels.

3.4. Mise à l'épreuve d'un des corpus du projet

Afin d'illustrer les performances d'un des corpus que j'ai créés, le corpus *Informations pratiques*, je présente ci-dessous la comparaison des résultats d'une requête dans ce corpus et dans trois corpus existants, le *Braun Corpus* (Divsic *et al.*, 2008), *DeWac* (Baroni, 2006) et le *English-German Translation Corpus*.

Tableau 2 - Nombre d'occurrences pertinentes pour "Eintritt" [entrée] dans quatre corpus

	Braun Corpus	***DeWac* (*Sketch Engine*)**	***English-German Translation Corpus***	***Informations pratiques***
Nombre total d'occurrences de "Eintritt"	38	46130 pour "Eintritt" 5849 pour "Eintritt + frei", mais il y a redondance	11	117
Occurrences pertinentes de "Eintritt"	11 après élimination des occurrences non pertinentes	au moins 5000 ; nécessité d'élimination des lignes redondantes ou non pertinentes	10 après élimination d'une occurrence non pertinente	14 après élimination des lignes redondantes

Pour obtenir des occurrences qui représentent un genre textuel donné, les lignes pertinentes du *Braun Corpus* doivent être présélectionnées. Le corpus *DeWac* permet de trouver une quantité impressionnante d'occurrences qui conviennent, mais à condition de choisir une combinaison de mots clés restreignant suffisamment le nombre de lignes, ce qui nécessite également une présélection. La collection de corpus *English-German Translation Corpus* a l'avantage de contenir un corpus spécialisé dans le domaine du tourisme. Ceci permet d'obtenir des résultats qui sont tous pertinents. Toutefois, les types de requêtes implémentés paraissent limités en 2010. Le corpus *Informations pratiques* permet de trouver un nombre satisfaisant d'occurrences de "Eintritt". Elles sont toutes pertinentes. Certes, il y a redondance, mais la redondance peut être porteuse de signification. Cette comparaison fournit des arguments en faveur la création de corpus réfléchis, même de très petite taille. Le paragraphe suivant décrira les étapes de création d'un tel corpus.

4. Création de corpus réfléchis

La création de corpus inclut les étapes suivantes, dont certaines sont facultatives. Après la recherche et l'exportation des données du corpus, il peut être utile d'enlever les marques de formatage dues à la publication en ligne. Afin de pouvoir utiliser certains logiciels, il est nécessaire d'enregistrer préalablement le texte sous un format particulier (format rtf, texte brut, etc.). Selon les besoins, on procèdera à la segmentation du corpus. Cela consiste à ajouter dans le fichier source des marques délimitantes. Pour certains projets, il convient de (faire) annoter les données, c'est-à-dire, d'ajouter de l'information linguistique ou autre aux unités du corpus.

Pour effectuer les gestes décrits ci-dessus, il est possible de se servir de collections d'outils dont l'emploi nécessite un degré d'expertise technique plus ou moins élevé. Dans ce qui suit, je me référerai à des outils gratuits et paraissant suffisamment "simples" pour une utilisation

dans un contexte d'enseignement-apprentissage des langues à un niveau peu avancé à intermédiaire.

4.1. Outils de recherche de données

Pour trouver les données pouvant alimenter un corpus réfléchi, j'ai utilisé un moteur de recherche permettant d'identifier des sites publiant des textes qui cadrent bien avec les genres textuels recherchés. J'aurais souhaité automatiser davantage la collecte des données, par exemple, en profitant de l'agrandissement automatique du corpus à l'aide d'un fil RSS. Toutefois, je n'ai pas trouvé d'outil donnant entièrement satisfaction à ce jour (pour un aperçu plus complet, voir Bower et Pearson, 2002 : 61-62). Pour ma recherche, j'ai choisi de sélectionner des textes publiés sur des sites de musées ou établissements culturels en ligne et de les copier dans un fichier *Word*. Ces textes sont protégés par un *copyright*. Suite à ma demande, ils ont gracieusement été mis à ma disposition par les propriétaires des sites concernés. Une autre option est la création de corpus à l'aide de l'outil *BootCat* (Baroni et Bernardini, 2003), proposé par *Sketch Engine*, qui rassemble des données publiées sur la Toile en fonction de mots clés spécifiés par le créateur de corpus. Les sites ainsi identifiés, contenant les mots clés, peuvent être cochés ou décochés. Ce procédé a des avantages, comme la rapidité, la quantité potentiellement suffisante de données rassemblées, l'option de recueillir uniquement des textes libres de droit, la possibilité de décider quelles données on souhaite garder ou éliminer et l'annotation automatique des données. Lors de la première phase de la recherche, j'ai créé un corpus du domaine du tourisme en me servant de cet outil. J'ai choisi des mots clés que j'ai estimés pertinents pour le domaine du tourisme, comme, par exemple "Tourismus, Eintritt, Kinder, Erwachsene, Behinderte, Führung" [Tourisme, entrée, enfants, adultes, handicapés, visite guidée]. Après une période d'essai, j'ai toutefois renoncé à son utilisation, car l'appartenance au genre touristique n'a pas été garantie pour une proportion suffisante d'occurrences. J'ai été contrainte de faire un travail important de présélection d'occurrences pertinentes. J'ai donc abandonné le travail sur ce corpus au cours de la recherche, tout en regrettant deux de ses plus grands atouts : les corpus créés avec *BootCat* sont annotés automatiquement (j'émets toutefois des réserves concernant la qualité du résultat pour l'allemand), et ils peuvent être interrogés à l'aide d'outils fournis par *Sketch Engine* dont l'interface convient pour un public d'apprenants de LE de niveau A2/B1.

4.2. Outils de segmentation

Il peut s'avérer utile de segmenter le corpus en plusieurs parties et sous-parties. Cela permet, entre autres, de comparer les caractéristiques linguistiques des sections d'un texte (cf. 2.3.). Ce procédé de délimitation peut concerner les éléments suivants : lexèmes, phrases, paragraphes, sections, textes d'un même auteur ou d'une même période de publication, etc. Les logiciels délimitent en général automatiquement les lexèmes d'un corpus. D'autres frontières doivent (ou peuvent) être marquées manuellement. À titre d'exemple, selon le manuel d'utilisation de *Lexico3* (Fleury *et al.*, 2003 : 10), la clé indiquant l'année de publication d'une partie segmentée peut être encodée comme suit : <Année=1793>. "Année" indique le type de la clé, et "1793" indique son contenu. Ce code précède la partie concernée.

4.3. Outils d'annotation

L'annotation des unités du corpus, donc l'ajout d'informations linguistiques ou autres, est requise si l'on souhaite faire des requêtes qui ne sont pas basées uniquement sur des chaînes de caractères. Parfois, on peut souhaiter, par exemple, associer un mot clé à un verbe conjugué, sans en nommer un en particulier. Cette étape de traitement du corpus demande certaines connaissances en TAL (Traitement Automatique des Langues). Des collections d'outils, comme *Sketch Engine* ou *Nooj*, proposent une annotation automatique des données du corpus. L'outil d'annotation en ligne *TreeTagger output visualisation module* (Koller, nd)

est particulièrement facile d'emploi. Pour obtenir de plus amples informations, il peut être utile de lire le paragraphe 3.2. de Silbersztein et Tutin (2005) et le chapitre 5 de Bowker et Pearson (2002).

4.4. Outils d'interrogation

Les corpus peuvent être explorés à l'aide d'outils de statistique textuelle. Le concordancier en est un exemple. Ce logiciel permet de rechercher et de regrouper toutes les occurrences d'un mot clé donné du corpus. Les résultats sont affichés sous forme de lignes contenant le mot clé, selon une mise en forme qui le rend particulièrement visible.

Tableau 3 – Lignes de concordance obtenues avec Sketch Engine

und eine kostenlose	**Führung**	. Jüdisches Museum
stündliche kostenlose	**Führungen**	durch die Schiller-Ausstellung
bis 18 Jahre frei,	**Führungen**	nach Voranmeldung

D'autres outils servent à créer des listes de mot, à repérer les mots clés principaux du corpus, à localiser visuellement les emplacements d'un mot clé dans une ou plusieurs de ses parties ou à trouver ses cooccurrents les plus fréquents, etc. Ces outils sont parfois regroupés dans un même logiciel. C'est le cas pour *AntConc3.2.1w.* (Anthony, 2007), *Lexico3* (Fleury *et al.*, 2003) et *Text-based Concordances (v. 2.1.)* de Cobb (2009). Ces collections d'outils donnent à l'utilisateur l'occasion de télécharger et d'explorer son propre corpus.

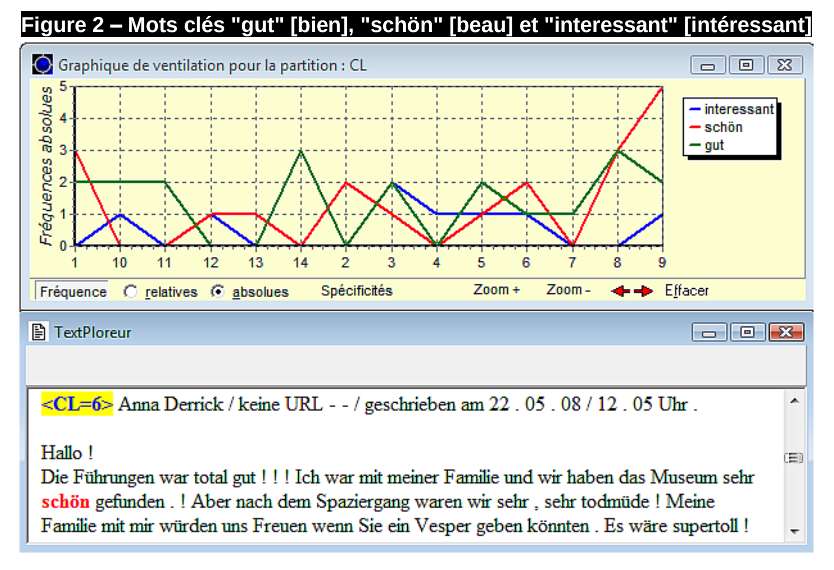
Figure 1 – Segments répétés

Lg	Segment	Frq
2	für die	15
2	ist nicht	10
2	das Museum	8
2	Ich finde	7
2	Ich habe	7
2	das ist	6
2	Hallo Ich	6
2	Danke schön	6
2	sehr schön	6

Dans ce qui suit, je présenterai deux fonctions de *Lexico3*, une collection d'outils conçue pour des spécialistes de la langue. *AntConc3.2.1w.* et *Text-based Concordances (v. 2.1.)* sont plus faciles à prendre en main que *Lexico3*. Toutefois, en les utilisant, j'ai rencontré des problèmes liés au formatage du texte allemand. De plus, *Lexico3* offre davantage de possibilités de segmentation et de représentation visuelle que les deux autres logiciels.

L'option "segments répétés", proposée par *Lexico3*, permet d'évaluer la fréquence, dans le corpus, de groupes de deux mots ou plus qui se suivent. Cela donne, entre autres, l'occasion d'observer si les apprenants ont eu recours à des collocations, comme l'illustre la figure 2. Y apparaissent des segments ayant un caractère figé, comme, par exemple, "danke schön" [merci bien].

Figure 2 – Mots clés "gut" [bien], "schön" [beau] et "interessant" [intéressant]



La figure 2 permet de voir comment la fréquence de mots clés d'un corpus peut être graphiquement représentée à l'aide de *Lexico3*. J'ai segmenté le corpus contenant les textes des apprenants qui représentent le genre "Commentaire dans le Livre d'or électronique d'un musée". L'indication <CL=6> que j'ai ajoutée manuellement avant le texte représenté en figure 2 signifie "début d'un texte de type 'Commentaire dans un Livre d'or' ; 6ème texte". Cette figure juxtapose le 6ème texte du corpus, celui de Camille, et un graphique contenant les mots clés "gut" [bien], "schön" [beau] et "interessant" [intéressant]. Cela permet de voir que dans cette partie du corpus, correspondant au texte de Camille, les trois mots clés sont représentés, ce qui n'est pas le cas dans le texte de Corentin (partie 14) dans lequel seul le mot clé "gut" apparaît. Il est également possible de comparer de cette façon le lexique d'un corpus de référence et d'un corpus d'apprenants (Granger *et al.*, 2002).

5. Pour finir

Le choix de considérer les corpus comme espaces discursifs fournit des arguments forts en faveur du recours aux corpus réfléchis pour l'enseignement-apprentissage des langues étrangères. Il est parfois nécessaire de constituer de telles ressources soi-même. Ce procédé est coûteux en temps, et il demande certaines compétences techniques. Cela peut justifier la collaboration entre linguistes, didacticiens et spécialistes en TAL, au moins pour des projets d'une certaine ampleur. De tels projets devraient viser à proposer des corpus offrant deux types d'interfaces : une interface pour les professionnels en linguistique de corpus et une interface gérable par des non-spécialistes en informatique, en proposant un nombre suffisant de types de requête.

Références bibliographiques

Tous les liens étaient actifs en avril 2010.

BOMMIER-PINCEMIN, B. (1999). *Diffusion ciblée automatique d'informations :conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne, 6 avril 1999. Extrait disponible sur http://www.revue-texto.net/Corpus/Publications/pincemin_ad_1999.pdf

BOWKER, L. & PEARSON, J. (2002). *Working with Specialized Language. A practical guide to using corpora.* London, New York : Routledge.

BRAUN, S. & CHAMBERS, A. (2006). Elektronische Ressourcen für den Fremdsprachenunterricht. In JUNG, U. (dir.). *Praktische Handreichung für Fremdsprachenlehrer*. Frankfurt/M : Peter Lang (troisième édition, revue et corrigée). pp. 330-337. Disponible sur : http://epubs.surrey.ac.uk/cgi/viewcontent.cgi?article=1006&context=translation

COBB, T. Corpus builder. http://www.lextutor.ca/tools/, (2009).

GRANGER, S., HUNG, J., & PETCH-TYSON, S. (dir.) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam, Philadelphia : John Benjamins Publishing Company.

HABERT, B. & NAZARENKO, A. & SALEM, A. (1997). *Les linguistiques de corpus.* Paris : Armand Colin.

KUPIETZ, M. & KEIBEL, H. (2008). *DeReKo. Das Archiv gegenwartssprachlicher Referenzkorpora am IDS Mannheim*. <philtag n=7> , journée d'étude. Trier, 13 octobre 2008.

LÜDELING, A. & WALTER, M. (2009). Korpuslinguistik für Deutsch als Fremdsprache – Sprachvermittlung und Spracherwerbsforschung. Version élargie de LÜDELING, A. & WALTER, M. (à paraître). Korpuslinguistik. In FANDRYCH, C., HUFEISEN, B., KRUMM, H.-J. & RIEMER, C. (dir.). *Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch.* 2ème édition, revue et corrigée (= HSK, 19). Berlin, New York : Mouton de Gruyter. 37 pages.

MAINGUENEAU, D. (2008). *L'énonciation en linguistique française*. Paris : Hachette. 2éme édition.

MAYAFFRE, D. (2005). Rôle et place des corpus en linguistique : réflexions introductives. *Texto!*, vol. 10, n° 4. http://www.revue-texto.net/1996-2007/Corpus/Publications/Mayaffre_Corpus.html

NARITA, M., KUROKAWA, K. & UTSURO, T. (2003). Case study on the development of a computer-based support tool for assisting Japanese software engineers with their English writing needs. Professional Communication, *IEEE Transactions on Publication*, vol. 46/3. pp. 194-209.

O'KEEFFE, A. & MCCARTHY, M. & CARTER, R. (2007). *From Corpus to Classroom. Language Use and Language Teaching.* Cambridge : Cambridge University Press.

RASTIER, F. (2004). Enjeux épistémologiques de la linguistique de corpus. *Texto !*, juin 2004. Rubrique Dits et inédits. http://www.revue-texto.net/1996-2007/Inedits/Rastier/Rastier_Enjeux.html

TEUBERT, W. (2006). Korpuslinguistik, Hermeneutik und die soziale Konstruktion der Wirklichkeit [La linguistique de corpus, l'herméneutique et la construction sociale de la réalité]. *Linguistik Online* 28, 3/06. pp. 41-60. http://www.linguistik-online.de/28_06/teubert.html

Corpus et logiciels

ANTHONY, T. (2007). *AntConc3.2.1w*. http://www.antlab.sci.waseda.ac.jp/antconc_index.html

BARONI, M. (2006). *DeWac.* Corpus interrogeable sur le site de *Sketch Engine* (Kilgarriff *et al.).*

BARONI, M. & BERNARDINI, S. (2003). *BootCaT toolkit 0.1.2. (Bootstrapping Corpora and Terms from the Web).* Outil de création de corpus. http://sslmit.unibo.it/~baroni/bootcat.html

CHEMNITZ ENGLISH-GERMAN TRANSLATION CORPUS. Université de Chemnitz. http://ell.phil.tu-chemnitz.de/search/

COBB, T. (2009b). *Text-Based Concordances (v. 2.1.)*. http://www.lextutor.ca/concordancers/text_concord/

COSMAS II (2010). Institut für Deutsche Sprache, Mannheim. http://www.ids-mannheim.de/cosmas2/web-app/

DIVSIC, D., ROTT, S. & HORST, M. (2008). *Braun Corpus.* Corpus interrogable à l'aide du concordancier *Konkordanzer-Corpus Deutsch (v5.1).* http://www.lextutor.ca/concordancers/concord_g.html

FLEURY, S., LAMALLE, C., MARTINEZ, W. & SALEM, A. (2003). *Lexico3*. *Outils de statistique textuelle*. Paris : Université de Paris 3. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/

KILGARRIFF, A., RYCHLY, P. & POMIKALEK, J. (nd). *Sketch Engine*. http://www.sketchengine.co.uk/

KOLLER, T. (nd). *TreeTagger output visualisation module.* http://vsac.cele.nottingham.ac.uk/~ccztk/treetagger.php

MANNHEIMER KORPUS 1 (2010). Institut für Deutsche Sprache, Mannheim. http://www.ids-mannheim.de/cosmas2/projekt/referenz/korpora1.html?sigle=MK1

SILBERZTEIN, M. &TUTIN, A. (2005). NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE. *Alsic (Apprentissage des Langues et Systèmes d'Information et de Communication)*, vol. 8, n° 2, 2005, mis en ligne le 15 décembre 2005. pp. 123-134. http://alsic.revues.org/index336.html